QÜESTIIÓ, vol. 21, 1 i 2, p. 109-149, 1997

ESTIMADORES DE RAZÓN: UNA REVISIÓN

JOSÉ A. MAYOR GALLEGO*
Universidad de Sevilla

En este trabajo de revisión exponemos los principios básicos de la utilización de información adicional en la estimación de parámetros definidos sobre poblaciones finitas, mediante el empleo de estimadores de razón. Aunque la introducción de este tipo de estimadores se realizó inicialmente desde un punto de vista «heurístico», basado en la existencia de relaciones de proporcionalidad directa «aproximada» entre la variable de estudio y otras variables más controladas, su estudio detallado se desarrolla a partir del enfoque de modelos de superpoblación. La problemática principal que presentan este tipo de estimaciones es la existencia de sesgo, lo que obliga a utilizar diseños muestrales específicos o a modificar las expresiones de los estimadores con el fin de obtener estrategias insesgadas o con sesgo reducido, pero manteniendo la simplicidad en la estimación del error de muestreo.

Ratio estimators: a review

Keywords: Muestreo, poblaciones finitas, estimadores de razón.

Clasificación AMS: 62DO5

*José A. Mayor Gallego. Dpto. Estadística e Investigación Operativa. Universidad de Sevilla. C/Tarfia s/n. 41012 SEVILLA.

– Article rebut el maig de 1996.

– Acceptat l'abril de 1997.

1. INTRODUCCIÓN

La utilización de información auxiliar es un recurso muy extendido en los diversos ámbitos del muestreo en poblaciones finitas, siendo su principal objetivo la obtención de estimaciones más acuradas.

En términos generales, podemos afirmar que la información auxiliar, generalmente suministrada por una o más **variables auxiliares**, conocidas o controladas al menos en cierto grado, puede ser aplicada en la fase de muestreo, en la fase de estimación o en ambas.

Así, los diseños ΠPS y PPS, es decir, con probabilidades de inclusión o de selección de elementos, proporcionales al tamaño, utilizan probabilidades de selección de elementos, que están afectadas por los valores de una variable auxiliar, X. También, en el muestreo estratificado, se emplean variables auxiliares en cuestiones tales como la afijación y la definición de estratos.

En este trabajo revisaremos diferentes formas de construir estimadores, con una estructura matemática de tipo **fraccional**, y que utilicen en la forma más adecuada, dicha información auxiliar, buscando la obtención de buenas estimaciones. Por supuesto, ello no va en detrimento de utilizar estos estimadores en combinación con diseños muestrales que también incorporen información auxiliar como los mencionados anteriormente. Una clasificación pormenorizada de éstas y otras formas de empleo de la información auxiliar puede verse en Hedayat y Sinha (1991).

Este tipo de estimadores de **razón** resultan muy apropiados cuando se presenta una relación aproximada de proporcionalidad directa entre la variable de estudio y otras variables auxiliares. Estas relaciones aparecen con frecuencia en situaciones reales. Por ejemplo, para estimar el contenido total de azúcares de un gran cargamento de naranjas, podemos utilizar la proporcionalidad existente entre el peso del fruto y el de azúcares que contiene, empleando el muestreo para estimar el factor de proporcionalidad. O para estimar el total de automóviles en una población, podemos tener en cuenta la proporcionalidad aproximada, existente entre éstos y el número de habitantes. De esta forma, la existencia de este tipo de relaciones puede ayudarnos a obtener estimaciones más precisas, aunque como veremos, con la contrapartida de la aparición de sesgos en las mismas.

Así, en la sección 2, iniciaremos el desarrollo de estas cuestiones estudiando lo que denominamos **soluciones heurísticas**, basadas en considerar el parámetro a estimar como una derivación de una expresión más compleja, sobre la cual se sustituyen ciertas cantidades poblacionales por muestrales. Un estudio posterior permitirá calibrar si hemos obtenido un estimador adecuado, y encontrar las mejores condiciones para su aplicación.

En la sección 3., enfocaremos nuestro estudio bajo la perspectiva de los modelos de superpoblación, desarrollando la solución general obtenida para el caso de diseño muestral aleatorio simple, $\mathrm{MAS}(N,n)$, y para el caso de diseños ΠPS, es decir, con probabilidades de inclusión proporcionales al tamaño.

En la sección 4., se estudian varias estrategias insesgadas, basadas en el diseño de Lahiri-Midzuno, y en el diseño $\mathrm{MAS}(N,n)$ combinado con estimadores especiales como el de Hartley-Ross y el de Mickey. Estas estrategias presentan estimadores de la varianza complicado, por lo que, en esta misma sección, introducimos varias estrategias cuasi-insesgadas, con estimadores de la varianza más simples.

La sección 5., se dedica al estudio del estimador de razón multivariante, y la sección 6., a las distintas formas de combinar el estimador de razón con la estructura de estratos.

Finalmente, en la sección 7. estudiamos propiedades de optimalidad del estimador de razón, bajo el modelo de superpoblación de proporcionalidad directa, exponiendo los resultados clásicos sobre optimalidad de ciertas estrategias basadas en muestreo intencional.

En lo que sigue, denotaremos por U a la población finita bajo estudio, siendo sus elementos,

$$U = \{1,2,3,\ldots,N\}$$

También denotaremos por Y una variable genérica, definida sobre U que asocia a cada elemento, $i \in U$, un número real, Y_i.

Supondremos que las estimaciones se realizan a partir de una muestra, m, obtenida de la población mediante un determinado diseño muestral. En particular, emplearemos frecuentemente el diseño muestral formado por todas las muestras posibles (en el sentido de subconjuntos) de tamaño n, con distribución de probabilidad uniforme sobre las mismas, al que denominamos **diseño muestral aleatorio simple**, ya mencionado previamente, y que denotaremos $\mathrm{MAS}(N,n)$.

2. SOLUCIONES HEURÍSTICAS

Si suponemos que el parámetro a estimar es la media poblacional de la variable Y,

$$\overline{Y} = \frac{1}{N}\sum_{i\in U} Y_i$$

y que X es una variable auxiliar perfectamente conocida para todos los elementos de U, podemos considerar la expresión,

$$\overline{Y} = \frac{\overline{Y}}{\overline{X}}\overline{X} = R\overline{X}$$

a partir de la cual podemos definir, **heurísticamente**, el estimador,

$$\widehat{\overline{Y}}_R = \widehat{R}\overline{X}$$

La forma final del estimador dependerá del diseño muestral utilizado. Por ejemplo, si la muestra se obtiene mediante diseño muestral aleatorio simple, $\mathrm{MAS}(N,n)$, podemos estimar R por la razón de medias muestrales, obteniendo,

$$\widehat{\overline{Y}}_R = \frac{\bar{y}}{\bar{x}}\overline{X}$$

Obviamente, la eficiencia de este estimador depende en gran medida de la relación existente entre las variables Y y X, siendo el caso más favorable aquel en el que existe una relación aproximada de proporcionalidad entre la variable de estudio y la variable auxiliar. Gráficamente ello significa una nube de puntos concentrada en las proximidades de una línea recta que pasa por el origen. Véase la Figura 1.

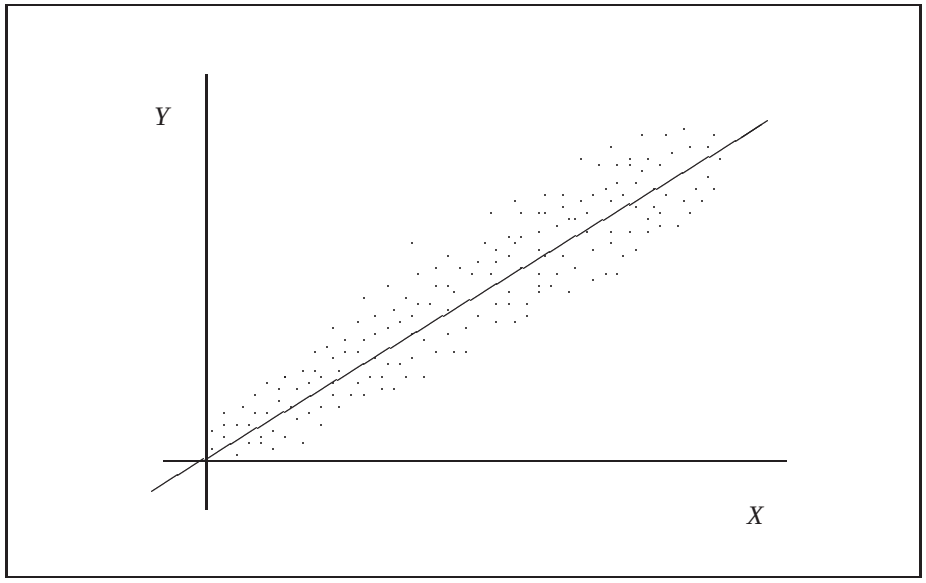

Figura 1. Relación de proporcionalidad aproximada entre dos variables.

El caso ideal, aunque utópico, se daría cuando la proporcionalidad entre las variables es exacta. En tal caso $V[\widehat{\overline{Y}}_R] = 0$, y la estimación coincide con el verdadero valor. En general, y suponiendo diseño MAS(N,n), basta aplicar las expresiones usuales para la varianza aproximada de la razón, y su estimación, véase Fernández y Mayor (1995), para obtener inmediatamente las conocidas expresiones,

$$\begin{aligned} V[\widehat{\overline{Y}}_R] &\approx \frac{1-f}{n}\left(S_y^2 + R^2 S_x^2 - 2RS_{xy}\right) \\ \widehat{V}[\widehat{\overline{Y}}_R] &= \frac{1-f}{n}\left(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}\right) \end{aligned}$$

donde S_y^2, S_x^2, S_{xy} denotan cuasivarianzas y cuasicovarianza poblacionales, y s_y^2, s_x^2, s_{xy} las correspondientes muestrales. También denotamos $f = n/N$.

Es interesante observar que la solución obtenida heurísticamente volverá a aparecer al aplicar el enfoque predictivo, por lo que pospondremos para más adelante un estudio en profundidad de esta solución, en lo que se refiere al tratamiento del sesgo y del error cuadrático medio.

También es necesario resaltar que, en el enfoque heurístico se está empleando, aunque no *a priori* o de una forma manifiesta, la existencia de una relación aproximada de proporcionalidad directa, entre las variables Y y X, y en este sentido, hemos de recordar los trabajos pioneros de algunos científicos, en el campo de la demografía, que han utilizado esta idea.

Así, son clásicos los estudios del inglés John Graunt sobre la estimación del número de habitantes de Londres. Sus resultados se publicaron en el famoso trabajo *Natural and Political Observations made upon the Bills of Mortality*, aparecido en 1662. En este estudio, Graunt investigó un conjunto de familias pertenecientes a determinadas parroquias de la ciudad de Londres, donde los registros resultaban fiables, y observó que había un promedio de tres fallecimientos anuales en 11 familias, siendo la cantidad total de fallecimientos por año en esta ciudad de aproximadamente de 13000. De este forma, Graunt concluyó que el número de familias era de 48000, y suponiendo un tamaño medio familiar de 8, estimó en 384000 el número de habitantes de la ciudad.

Como puede observarse, en estos cálculos está implícito un modelo de proporcionalidad entre el número de fallecimientos y el de familias y también entre el número de éstas y el de habitantes. También hay que notar como Graunt no realizó ningún estudio adicional encaminado a cuantificar los posibles errores cometidos. Véanse Chang (1976) y Hald (1990) para un estudio pormenorizado del trabajo de Graunt.

Otro precedente histórico de gran relevancia lo constituyen los estudios sobre la población de Francia, llevados a cabo por Laplace, y cuyos primeros resultados

aparecieron en 1786. Sus métodos de muestreo y estimación fueron similares a los empleados por Graunt, pero el científico francés vio la necesidad de tener en cuenta de alguna forma la precisión de los resultados obtenidos, tanto en su control, seleccionando una muestra de calidad, como en su medición.

A partir de una muestra, Laplace estimó la población total del país utilizando una estimación de razón, empleando los nacimientos ocurridos el año precente como variable auxiliar. Adicionalmente, calculó la distribución de la diferencia entre el verdadero valor y el estimado, aproximando esta distribución por una normal. Sus métodos y resultados aparecieron en la clásica obra *Théorie Analytique des Probabilités*, publicada en 1812. Véase Cochran (1978) y Chang (1976).

3. MODELO DE SUPERPOBLACIÓN DE PROPORCIONALIDAD DIRECTA

El enfoque considerado en el apartado anterior, que denominamos **heurístico**, en cierto modo se contrapone, metodológicamente, al que vamos a considerar ahora, más formal, y denominado **enfoque predictivo**. Este enfoque se basa en suponer un «modelo» o relación funcional entre la variable de estudio, Y, y la variable auxiliar, X, de la forma $Y = f(X)$. Si $f(\cdot)$ fuera conocida completamente, el conocimiento de X nos llevaría al de Y y por tanto al de cualquier parámetro $\theta(Y)$.

Usualmente, $f(\cdot)$ es desconocida y su determinación sólo puede realizarse de un modo **aproximado**, a partir del conocimiento de la variable X, y de la información suministrada por el estadístico,

$$\{(X_i, Y_i) \,|\, i \in m\}$$

Con dicha información, buscaremos la función $\widehat{f}(\cdot)$ que «mejor» explique la relación observada y que denominaremos **función predictora**.

En este sentido, es muy importante realizar estudios exploratorios de los datos muestrales, por ejemplo dibujando la **nube de puntos**, que proporcionen indicios sobre las pautas que relacionan X con Y. Véase Fernández y Mayor (1995).

El siguiente paso será estimar $\theta(Y)$ mediante $\theta(\widehat{Y})$, siendo $\widehat{Y}_i$, $i \in U$ los valores aproximados a los verdaderos valores Y_i, $i \in U$, proporcionados por la función predictora $\widehat{f}(\cdot)$, es decir,

$$\widehat{Y}_i = \widehat{f}(X_i) \qquad i \in U$$

Por ejemplo, para un parámetro lineal del tipo $\theta(Y) = \sum_{i \in U} a_i Y_i$ se tendrá,

$$\theta(Y) = \sum_{i \in U} a_i Y_i = \sum_{i \in U} a_i \widehat{Y}_i + \sum_{i \in U} a_i (Y_i - \widehat{Y}_i)$$

cuyo primer sumando es conocido, y el segundo es desconocido, siendo función del error de estimación de $f(\cdot)$. Si es posible despreciar este segundo sumando, entonces podemos dar como estimador,

$$\widehat{\theta}_1(Y) = \sum_{i \in U} a_i \widehat{Y}_i$$

en otro caso, podemos estimarlo, por ejemplo mediante el estimador de Horvitz-Thompson, obteniendo el estimador alternativo,

$$\widehat{\theta}_2 = \sum_{i \in U} a_i \widehat{Y}_i + \sum_{i \in m} a_i \frac{Y_i - \widehat{Y}_i}{\pi_i}$$

Es importante observar que en el **enfoque predictivo** se combinan dos procesos estadísticos, el **ajuste** y la **estimación**, dependiendo la bondad de las estimaciones de numerosos factores entre los que destacamos la habilidad en la conjunción de ambos procesos, la bondad del ajuste realizado y la estructura de la población en lo que atañe a las variables involucradas.

Con el fin de obtener el estimador de razón a partir del enfoque predictivo, vamos a suponer que la población, en relación a la variable de estudio, Y, y la variable auxiliar, X, posee el siguiente modelo de superpoblación, de **proporcionalidad directa**,

$$\begin{aligned} &Y_i = \beta X_i + \varepsilon_i \\ \\ &E_s[\varepsilon_i] = 0 \\ &V_s[\varepsilon_i] = \sigma^2 v(X_i) \\ &E_s[\varepsilon_i \varepsilon_j] = 0, \ i \neq j \end{aligned}$$

siendo $v(\cdot)$ una función conocida que marca la estructura de la varianza. Adicionalmente, supondremos que X toma únicamente valores no negativos.

Notemos que si fuera posible observar la totalidad de los valores $\{(X_i, Y_i) \,|\, i \in U\}$, como en el caso de un censo, podríamos obtener una estimación de β basándonos en el teorema de Gauss-Markov generalizado dado por C.R. Rao (1965), mediante la resolución del siguiente problema de minimización,

$$\min_{\beta} \quad \sum_{i \in U} \frac{(Y_i - \beta X_i)^2}{\sigma^2 v(X_i)}$$

lo que proporcionaría,

$$\widehat{\beta} = \frac{\sum_{i\in U} Y_i X_i / v(X_i)}{\sum_{i\in U} X_i^2 / v(X_i)}$$

Pero como sólo disponemos de la información suministrada por la muestra, utilizaremos el método de estimación propuesto por Kish y Frankel (1974) y Fuller (1975), consistente en reemplazar la suma poblacional por una estimación muestral, y más concretamente, si es la de Horvitz-Thompson, resolviendo el problema,

$$\min_{\beta} \quad \sum_{i\in m} \frac{(Y_i - \beta X_i)^2}{\sigma^2 v(X_i)\pi_i}$$

Para simplificar la solución del mismo, tomaremos $v(x) = x$, lo que representa una situación muy general, en la cual la varianza en la superpoblación aumenta proporcionalmente al valor de la variable auxiliar (que ha de ser no negativa). Con esta hipótesis, obtenemos, sin más que derivar, la siguiente ecuación normal,

$$\sum_{i\in m} \frac{Y_i - \widehat{\beta} X_i}{\pi_i} = 0$$

y la siguiente estimación de β,

$$\widehat{\beta} = \frac{\sum_{i\in m} Y_i/\pi_i}{\sum_{i\in m} X_i/\pi_i}$$

Si ahora empleamos el estimador $\widehat{\theta}_2$ para estimar la media poblacional, obtendremos,

$$\widehat{\theta}_2 \quad = \sum_{i\in U} \widehat{Y}_i/N + \sum_{i\in m} \frac{Y_i - \widehat{Y}_i}{N\pi_i} = \sum_{i\in U} \widehat{\beta} X_i/N + \sum_{i\in m} \frac{Y_i - \widehat{\beta} X_i}{N\pi_i} = \widehat{\beta}\overline{X} \stackrel{\triangle}{=} \widehat{\overline{Y}}_R$$

donde el segundo sumando se ha anulado por la ecuación normal. Así pues, hemos obtenido el siguiente estimador de la media poblacional,

$$\widehat{\overline{Y}}_R = \frac{\sum_{i\in m} Y_i/\pi_i}{\sum_{i\in m} X_i/\pi_i}\overline{X}$$

caso particular del de Sánchez-Crespo (1980).

El estudio de la varianza de $\widehat{\overline{Y}}_R$ dependerá del diseño muestral que se emplee, y en general, se puede realizar por los métodos usuales basados en aproximación lineal.

A continuación particularizaremos la solución obtenida, para el diseño muestral aleatorio simple, MAS(N,n), y para los diseños ΠPS, esto es, con probabilidades de inclusión de los elementos proporcionales a su valor de la variable auxiliar.

3.1. Diseño MAS(N, n)

Al ser las probabilidades de inclusión de primer orden para este diseño muestral, $\pi_i = n/N,\ \forall i \in U$, obtenemos el estimador,

$$\widehat{\overline{Y}}_R = \frac{\overline{y}}{\overline{x}}\overline{X}$$

que coincide con el obtenido heurísticamente, y obviamente es, en general sesgado, ya que,

$$B[\widehat{\overline{Y}}_R] = E[\widehat{\overline{Y}}_R] - \overline{Y} = \overline{X}E\left[\frac{\overline{y}}{\overline{x}}\right] - E[\overline{y}] = E[\overline{x}]E\left[\frac{\overline{y}}{\overline{x}}\right] - E[\overline{y}] = -\operatorname{Cov}\left[\overline{x}, \frac{\overline{y}}{\overline{x}}\right]$$

Con objeto de realizar un estudio cuantitativo de este sesgo, así como del error cuadrático medio y de la varianza, construiremos los valores,

$$\delta\overline{y} = \frac{\overline{y} - \overline{Y}}{\overline{Y}} \qquad \delta\overline{x} = \frac{\overline{x} - \overline{X}}{\overline{X}}$$

que nos servirán, adoptando la línea expuesta en David y Sukhatme (1974), para definir las siguientes cantidades,

$$\begin{aligned} B_k[\widehat{\overline{Y}}_R] &= \overline{Y}E\left[\sum_{i=0}^{2k-1}(-\delta\overline{x})^i(\delta\overline{y} - \delta\overline{x})\right] \\ \mathrm{ECM}_k[\widehat{\overline{Y}}_R] &= \overline{Y}^2 E\left[(\delta\overline{y} - \delta\overline{x})^2 \sum_{i=0}^{2k-2}(i+1)(-\delta\overline{x})^i\right] \end{aligned}$$

con $k \geq 1$ entero. Estas cantidades serán utilizadas como aproximaciones del sesgo y del error cuadrático medio, y con respecto a sus órdenes de aproximación se verifica el resultado que exponemos a continuación.

Teorema 1 *Bajo diseño muestral MAS(N,n), y suponiendo que la media muestral de X verifica $\overline{x} \geq x_0 > 0$, se tiene para las cantidades $B_k[\widehat{\overline{Y}}_R]$ y $\mathrm{ECM}_k[\widehat{\overline{Y}}_R]$,*

$$\left|B[\widehat{\overline{Y}}_R] - B_k[\widehat{\overline{Y}}_R]\right| \qquad \leq O(n^{-(k+1)})$$

$$\left|\mathrm{ECM}[\widehat{\overline{Y}}_R] - \mathrm{ECM}_k[\widehat{\overline{Y}}_R]\right| \qquad \leq O(n^{-(k+1)})$$

Para un estudio pormenorizado de este resultado, y su demostración, véanse David y Sukhatme (1974) y Sukhatme *et al.* (1984).

Tomando ahora $k = 1$, y teniendo en cuenta que entre el error cuadrático medio y la varianza existe la relación,

$$\mathrm{ECM}[\widehat{\overline{Y}}_R] = V[\widehat{\overline{Y}}_R] + \left(B[\widehat{\overline{Y}}_R]\right)^2$$

se obtienen fácilmente las siguientes aproximaciones.

Teorema 2

$$\begin{aligned}
B[\widehat{\overline{Y}}_R] &= \frac{1-f}{n}\left(\frac{\overline{Y}}{\overline{X}^2}S_x^2 - \frac{1}{\overline{X}}S_{xy}\right) + O(n^{-2}) = O(n^{-1}) \\
\mathrm{ECM}[\widehat{\overline{Y}}_R] &= \frac{1-f}{n}\left(S_y^2 + \frac{\overline{Y}^2}{\overline{X}^2}S_x^2 - 2\frac{\overline{Y}}{\overline{X}}S_{xy}\right) + O(n^{-2}) = O(n^{-1}) \\
V[\widehat{\overline{Y}}_R] &= \frac{1-f}{n}\left(S_y^2 + \frac{\overline{Y}^2}{\overline{X}^2}S_x^2 - 2\frac{\overline{Y}}{\overline{X}}S_{xy}\right) + O(n^{-2}) = O(n^{-1})
\end{aligned}$$

El resultado anterior es importante por varias razones. Por una parte, nos dice que el sesgo puede ser reducido incrementando el tamaño muestral.

Por otra parte, proporciona una expresión del sesgo, aproximada hasta el orden $O(n^{-2})$. Similares consideraciones se derivan para el error cuadrático medio y la varianza.

Como una aplicación interesante, hemos realizado una comprobación empírica del comportamiento del sesgo, utilizando una población construida artificialmente, EXP1000, con $N = 1000$ elementos, para los cuales las variables Y y X están ligadas por la relación,

$$Y_i = 1000 + 10 \times X_i + 3 \times \varepsilon_i, \quad i = 1, \ldots, 1000$$

Los valores de X_i han sido generados de una distribución exponencial de media $\mu = 200$ y los de ε_i se han obtenido restando 50 a los valores generados a partir de una distribución exponencial de media 50.

Para cada uno de los valores $n = 10, 15, \ldots, 100$ hemos simulado 500 veces un muestreo aleatorio simple, MAS$(1000, n)$, y para cada valor de n hemos tabulado el **término guía** del sesgo,

$$B_1 = B_1[\widehat{\overline{Y}}_R] = \frac{1-f}{n}\left(\frac{\overline{Y}}{\overline{X}^2}S_x^2 - \frac{1}{\overline{X}}S_{xy}\right)$$

así como $\widehat{B} = \widehat{B}[\widehat{\overline{Y}}_R]$ calculado promediando $\widehat{\overline{Y}}_R - \overline{Y}$ sobre las 500 simulaciones. También hemos tabulado la diferencia, en valor absoluto, entre ambas cantidades. De esta forma hemos obtenido los resultados siguientes,

n	B_1	$\widehat{B}$	$\lvert\widehat{B} - B_1\rvert$
10	94.367097	79.384970	14.982127
15	62.593663	67.305330	4.711667
20	46.706946	50.874577	4.167631
25	37.174917	43.594316	6.419399
30	30.820230	26.349465	4.470765
35	26.281168	22.712932	3.568236
40	22.876871	20.871180	2.005691
45	20.229085	20.171039	0.058046
50	18.110856	12.119815	5.991041
55	16.377760	20.709838	4.332078
60	14.933513	18.269774	3.336261
65	13.711458	18.450887	4.739429
70	12.663982	14.513552	1.849570
75	11.756170	13.513885	1.757715
80	10.961834	12.005794	1.043960
85	10.260949	11.873348	1.612399
90	9.637941	10.489407	0.851466
95	9.080512	10.106495	1.025983
100	8.578827	9.283448	0.704621

Como puede verse, el comportamiento del sesgo estimado coincide con la expresión teórica, en lo que se refiere a su dependencia del tamaño muestral, n.

También la rápida disminución observada para $|\widehat{B}[\widehat{\overline{Y}}_R] - B_1[\widehat{\overline{Y}}_R]|$ parece concordar con el orden $O(n^{-2})$ hallado teóricamente. Una gráfica de estos valores se muestra en la Figura 2.

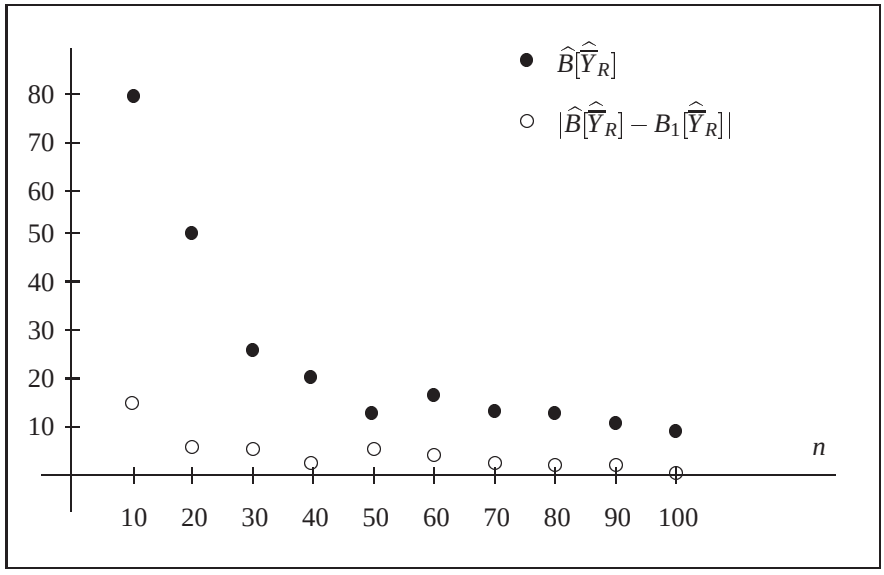

Figura 2. Sesgo estimado y su diferencia con el término orden $O(n^{-1})$.

Con respecto a la comparación entre el estimador de razón, $\widehat{\overline{Y}}_R$ y el estimador $\widehat{\overline{Y}} = \overline{y}$, que no emplea información auxiliar, se tiene,

Teorema 3 *Si el tamaño muestral es lo suficientemente grande para despreciar los términos de orden $O(n^{-2})$, entonces el estimador de razón, $\widehat{\overline{Y}}_R$ es más eficiente que $\widehat{\overline{Y}} = \overline{y}$ si el coeficiente de correlación lineal, ρ verifica,*

$$\rho > \frac{1}{2}\frac{\mathrm{CV}_x}{\mathrm{CV}_y}$$

donde $\mathrm{CV}_y = S_y/\overline{Y}$ denota el cuasicoeficiente de variación de Y, y análogo para X.

Resultado que se obtiene inmediatamente sin más que resolver en ρ la inecuación,

$$\frac{1-f}{n}\left(S_y^2+\frac{\overline{Y}^2}{\overline{X}^2}S_x^2-2\frac{\overline{Y}}{\overline{X}}S_{xy}\right)<\frac{1-f}{n}S_y^2$$

Observemos que la anterior condición para ρ, suele verificarse cuando existe una elevada correlación entre las variables, con lo cual obtendremos mejores resultados con el estimador de razón que con la media muestral simple. Por supuesto, ello presenta el inconveniente del sesgo, aunque este puede ser reducido por algunos procedimientos como el aumento del tamaño muestral, la aplicación de técnicas de tipo jackknife o la estimación del sesgo, siendo referencias fundamentales para el estudio de estas cuestiones, además de las citadas, Hansen, Hurwitz y Madow (1953), Cochran (1993) y Hedayat y Sinha (1991).

También es interesante la comparación con el estimador de regresión de la media,

$$\widehat{\overline{Y}}_{RG}=\bar{y}+\frac{s_{xy}}{s_x^2}(\overline{X}-\bar{x})$$

En este sentido, se verifica que si, como en el anterior resultado, despreciamos los términos de orden $O(n^{-2})$, entonces siempre es más eficiente el estimador de regresión que el de razón, es decir, $V[\widehat{\overline{Y}}_{RG}]\leq V[\widehat{\overline{Y}}_R]$. Véase Hedayat y Sinha (1991).

Finalmente añadiremos que tanto el sesgo, como el error cuadrático medio y la varianza de $\widehat{\overline{Y}}_R$, pueden ser estimados mediante,

$$\begin{aligned}
\widehat{B}[\widehat{\overline{Y}}_R] &= \frac{1-f}{n}\left(\frac{\bar{y}}{\bar{x}^2}s_x^2-\frac{1}{\bar{x}}s_{xy}\right)\\
\widehat{\mathrm{ECM}}[\widehat{\overline{Y}}_R] &= \frac{1-f}{n}\left(s_y^2+\frac{\bar{y}^2}{\bar{x}^2}s_x^2-2\frac{\bar{y}}{\bar{x}}s_{xy}\right)\\
\widehat{V}[\widehat{\overline{Y}}_R] &= \frac{1-f}{n}\left(s_y^2+\frac{\bar{y}^2}{\bar{x}^2}s_x^2-2\frac{\bar{y}}{\bar{x}}s_{xy}\right)
\end{aligned}$$

3.2. Diseños ΠPS

Para estos diseños, suponiendo tamaño de muestra fijo, n, se tiene $\pi_i=nX_i/T(X)$, con lo cual, el estimador de razón adopta la forma,

$$\widehat{\overline{Y}}_R=\frac{1}{n}\sum_{i\in m}\frac{Y_i}{X_i}\overline{X}$$

Es interesante observar que dicho estimador coincide con el estimador de Horvitz-Thompson de la media, empleando X como variable auxiliar y probabilidades de inclusión proporcionales al tamaño, siendo pues la estimación insesgada. El estudio de su varianza se realizará por los métodos usuales. Por ejemplo, aplicando las expresiones de Yates-Grundy-Sen, obtendremos,

$$V[\widehat{\overline{Y}}_R] = -\frac{\overline{X}^2}{2n^2}\sum_{i,j\in U}(\pi_{ij}-\pi_i\pi_j)\left(\frac{Y_i}{X_i}-\frac{Y_j}{X_j}\right)^2$$

$$\widehat{V}[\widehat{\overline{Y}}_R] = -\frac{\overline{X}^2}{2n^2}\sum_{i,j\in m}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{Y_i}{X_i}-\frac{Y_j}{X_j}\right)^2$$

Así, la varianza disminuye cuanto más ajustada es la relación de proporcionalidad entre las variables Y y X.

4. ESTRATEGIAS INSESGADAS Y CUASI-INSESGADAS

En este apartado estudiaremos algunas combinaciones de diseño y estimador que proporcionan estimaciones insesgadas o cuasi-insesgadas, es decir, con sesgo de orden $O(n^{-2})$. Estas estrategias se basan, tanto en las particularidades del diseño muestral, como en modificaciones introducidas en la expresión del estimador.

4.1. Estrategia insesgada basada en el esquema de Lahiri-Midzuno

Lahiri (1951) y Midzuno (1952), independientemente, han descrito el esquema de muestreo consistente en la selección de un elemento, i, con probabilidad proporcional a su tamaño, es decir, $p_i = X_i/T(X),\ \forall i \in U$; y la selección de $n-1$ elementos adicionales mediante diseño $\mathrm{MAS}(N-1,n-1)$ en $U-\{i\}$. La importancia de este esquema radica en el siguiente resultado,

Teorema 4 *Bajo el diseño muestral originado por el esquema de Lahiri-Midzuno se verifica que* $\widehat{\overline{Y}}_R = (\overline{y}/\overline{x})\overline{X}$ *es un estimador insesgado de* $\overline{Y}$.

Demostración: Estudiemos en primer lugar el diseño muestral resultante. Su espacio muestral está formado por todas las muestras de tamaño n. Para calcular la probabilidad, $p(m)$, de una muestra del diseño, m, consideremos $m^* = (j_1, j_2, \ldots, j_n)$, muestra ordenada con los mismos elementos de m. Se tiene entonces,

$$p(m^*) = \frac{X_{j_1}}{T(X)}\frac{1}{(n-1)!\binom{N-1}{n-1}}$$

y para la muestra m será,

$$p(m) = \sum_{m^* \sim m} p(m^*) = \frac{\sum_{i \in m} X_i}{T(X)} \frac{1}{\binom{N-1}{n-1}}$$

donde la notación $m^* \sim m$ expresa que la suma se extiende a todas las muestras ordenadas, con los mismos elementos que la muestra m. Calculemos ahora la esperanza de $\widehat{\overline{Y}}_R$,

$$\begin{aligned} E[(\overline{y}/\overline{x})\overline{X}] &= \sum_{m \in M} p(m)(\overline{y}(m)/\overline{x}(m))\overline{X} = \frac{1}{T(X)} \frac{1}{\binom{N-1}{n-1}} \sum_{m \in M} n\overline{x}(m)(\overline{y}(m)/\overline{x}(m))\overline{X} \\ &= \frac{1}{N} \frac{1}{\binom{N-1}{n-1}} \sum_{m \in M} \sum_{i \in m} Y_i = \frac{1}{N} \frac{1}{\binom{N-1}{n-1}} T(Y) \binom{N-1}{n-1} = \overline{Y} \end{aligned}$$

■

Para estudiar la varianza de esta estimación, se pueden seguir varios caminos. Uno de ellos se basa en la técnica de linealización, tomando los términos lineales del desarrollo de Taylor de $\widehat{\overline{Y}}_R = (\overline{y}/\overline{x})\overline{X}$ en un entorno de $(\overline{Y}, \overline{X})$,

$$\widehat{\overline{Y}}_R = \frac{\overline{y}}{\overline{x}}\overline{X} \approx \overline{Y} + (\overline{y} - \overline{Y}) - \frac{\overline{Y}}{\overline{X}}(\overline{x} - \overline{X}) = \overline{Y} + (\overline{y} - R\overline{x}) = \overline{Y} + \overline{z}$$

donde $Z_i = Y_i - RX_i,\ i \in m$. Se tiene pues,

$$\begin{aligned} V[\widehat{\overline{Y}}_R] &\approx V[\overline{z}] = -\frac{1}{2n^2} \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_j)(Z_i - Z_j)^2 \\ &= \sum_{i,j \in U} c_{ij}(Z_i - Z_j)^2 \end{aligned}$$

siendo,

$$c_{ij} = -\frac{1}{2n^2}(\pi_{ij} - \pi_i \pi_j)$$

Y si tenemos en cuenta que para el diseño que estamos considerando se verifica, véase Fernández y Mayor (1995),

$$\begin{aligned} \pi_i &= \frac{N-n}{N-1} \frac{X_i}{T(X)} + \frac{n-1}{N-1} \\ \pi_{ij} &= \frac{n-1}{N-1} \left[\frac{N-n}{N-2} \frac{(X_i + X_j)}{T(X)} + \frac{n-2}{N-2} \right] \quad i \neq j \end{aligned}$$

sustituyendo y desarrollando, se obtiene,

$$c_{ij} = \frac{N-n}{2n^2(N-1)^2}\left[(N-n)\frac{X_iX_j}{(T(X))^2} + \frac{(n-1)(1-X_i/T(X)-X_j/T(X))}{N-2}\right] \quad i \neq j$$

De forma similar, empleando la expresión de Yates-Grundy-Sen para la varianza estimada, se puede obtener una estimación de $V[\widehat{\overline{Y}}_R]$.

Otra línea diferente para el estudio de la varianza, sin emplear técnicas aproximadas, es la introducida por Rao y Vijayan (1977). Estos autores consideran la siguiente forma cuadrática para expresar la varianza, obtenida mediante un cálculo directo,

$$V[\widehat{\overline{Y}}_R] = \sum_{i\in U}\alpha_{ii}Y_i^2 + \sum_{\substack{i,j\in U\\ i\neq j}}\sum\alpha_{ij}Y_iY_j$$

donde,

$$\begin{aligned}\alpha_{ii} &= \frac{\overline{X}}{n^2\binom{N}{n}}\left(\sum_{m\ni i}\frac{1}{\overline{x}(m)}\right) - \frac{1}{N^2} \qquad \forall\, i \in U\\ \alpha_{ij} &= \frac{\overline{X}}{n^2\binom{N}{n}}\left(\sum_{m\ni i,j}\frac{1}{\overline{x}(m)}\right) - \frac{1}{N^2} \qquad \forall\, i \neq j \in U\end{aligned}$$

Y aplicando los resultados clásicos sobre estimación insesgada de formas cuadráticas (véase por ejemplo Hedayat y Sinha (1991)), obtenemos directamente el siguiente teorema,

Teorema 5 *Todo estimador insesgado y no negativo de la anterior varianza, $V[\widehat{\overline{Y}}_R]$, adopta necesariamente la forma,*

$$\widehat{V}[\widehat{\overline{Y}}_R] = -\frac{1}{2}\sum_{i,j\in m}\sum\alpha_{ij}(m)X_iX_j\left(\frac{Y_i}{X_i}-\frac{Y_j}{X_j}\right)^2$$

verificando los coeficientes la condición de insesgadez,

$$\sum_{m\ni i,j}\alpha_{ij}(m)\,p(m) = \alpha_{ij} \qquad \forall\, i \neq j \in U$$

Como posibles elecciones de $\alpha_{ij}(m)$, Rao y Vijayan (1977) sugieren las siguientes,

(a) $\alpha_{ij}(m) = \alpha_{ij}/\pi_{ij} \qquad i,j \in m$

(b) $\alpha_{ij}(m) = \dfrac{\overline{X}^2}{n^2\overline{x}^2(m)} - \dfrac{\overline{X}(N-1)}{\overline{x}(m)Nn(n-1)} \qquad i,j \in m$

En Rao (1966, 1972), Lanke (1974), Rao y Vijayan (1977) y Hedayat y Sinha (1991) pueden encontrarse resultados adicionales relacionados con esta línea.

4.2. Estrategia insesgada basada en el estimador de Hartley-Ross

Esta estrategia utiliza como diseño muestral el aleatorio simple, $\mathrm{MAS}(N,n)$, en combinación con un estimador especial definido por Hartley y Ross (1954). Para construir este estimador, consideremos previamente el siguiente estimador, **heurístico**, de la media poblacional,

$$\widehat{\overline{Y}}'_R = \frac{1}{n}\sum_{i\in m}\frac{Y_i}{X_i}\overline{X} = \overline{z}\overline{X}$$

donde $Z_i = Y_i/X_i$. Este estimador es sesgado ya que,

$$\begin{aligned} B[\widehat{\overline{Y}}'_R] &= E[\overline{z}\overline{X}] - \overline{Y} = \overline{Z}\overline{X} - \overline{Y} \\ &= \overline{Z}\overline{X} - \overline{ZX} = -\frac{N-1}{N}S_{zx} \end{aligned}$$

siendo entonces $[-(N-1)s_{zx}/N]$ un estimador insesgado de dicho sesgo.

Podemos entonces, sin más que restar la estimación insesgada del sesgo, construir el siguiente estimador insesgado de **Hartley y Ross** para la media poblacional,

$$\begin{aligned} \widehat{\overline{Y}}_{HR} &= \overline{z}\overline{X} + \frac{N-1}{N}s_{zx} \\ &= \overline{z}\overline{X} + \frac{n(N-1)}{N(n-1)}(\overline{y} - \overline{zx}) \end{aligned}$$

La varianza de este estimador, así como su estimación, adoptan formas muy complicadas, lo que explica que el estimador de Hartley-Ross no se haya popularizado. Para su obtención, Robson (1957) ha empleado el formalismo de «polykays» multivariantes, obteniendo una expresión de $V[\widehat{\overline{Y}}_{HR}]$ en función de medias simétricas poblacionales, a partir de la cual, podemos obtener un estimador insesgado sin más que sustituir éstas por las correspondientes medias simétricas muestrales.

Es posible realizar una simplificación suponiendo que la población es infinita, en cuyo caso se obtiene, empleando la notación usual,

$$\lim_{N\to\infty} V[\widehat{\overline{Y}}_{HR}] = \frac{1}{n}\left(\sigma_y^2 + \overline{Z}^2\sigma_x^2 - 2\overline{Z}\sigma_{xy}\right) + \frac{1}{n(n-1)}\left(\sigma_z^2\sigma_x^2 + \sigma_{xz}^2\right)$$
$$\approx \frac{1}{n}\left(\sigma_y^2 + \overline{Z}^2\sigma_x^2 - 2\overline{Z}\sigma_{xy}\right)$$

aproximación obtenida también, independientemente, por Goodman y Hartley (1958). Estos autores proporcionan además el siguiente resultado acerca de la comparación del estimador de Hartley y Ross y el estimador de razón usual, $\widehat{\overline{Y}}_R = (\overline{y}/\overline{x})\overline{X}$,

Teorema 6 *Si se ignoran los términos de orden $O(n^{-2})$ y superior, el estimador $\widehat{\overline{Y}}_{HR}$ es más eficiente que el estimador $\widehat{\overline{Y}}_R = (\overline{y}/\overline{x})\overline{X}$ si y sólo si el coeficiente de regresión, β, entre las variables Y y X está más próximo a $\overline{Z}$ que $R = \overline{Y}/\overline{X}$. En caso de ser $\overline{Z} = R$, ambos estimadores son igualmente eficientes.*

Observemos que la condición propuesta en este teorema no es fácil de verificar en la práctica, véase Sukhatme *et al.* (1984). Otras aportaciones relacionadas pueden también consultarse en Ruiz y Santos (1989), y en Sahoo y Ruiz (1994).

4.3. Estrategia insesgada basada en el estimador de Mickey

Una metodología diferente, propuesta por Mickey (1959), se fundamenta en la partición al azar de una muestra aleatoria simple, *m*, de tamaño *n*, en *k* grupos, cada uno con $l = n/k$ elementos. Denotando por $m_1, m_2, \ldots, m_k$ a dichos grupos, se definen entonces las cantidades,

$$\overline{y}(m - m_j), \qquad \overline{x}(m - m_j) \qquad\qquad j = 1, \ldots, k$$

es decir, las medias de *Y* y *X* calculadas sobre las submuestras obtenidas omitiendo en *m*, sucesivamente, los grupos $m_1, m_2, \ldots, m_k$.

A partir de las cantidades anteriores, definimos,

$$\overline{y}_R^{(j)} = \frac{\overline{y}(m - m_j)}{\overline{x}(m - m_j)}\overline{X} \qquad j = 1, \ldots, k$$

$$\overline{y}_M^{(j)} = \overline{y}_R^{(j)} + \frac{k(N-n+l)}{N}\left[\overline{y} - \overline{y}_R^{(j)}\frac{\overline{x}}{\overline{X}}\right]$$

que nos sirven para construir el estimador de Mickey,

$$\widehat{\overline{Y}}_M = \frac{1}{k}\sum_{j=1}^{k} \overline{y}_M^{(j)}$$

cuya importancia radica en el siguiente resultado,

Teorema 7 *$\widehat{\overline{Y}}_M$ es un estimador insesgado de la media poblacional.*

Demostración: Para probarlo, es suficiente demostrar que para cada j, $\overline{y}_M^{(j)}$ es insesgado respecto a $\overline{Y}$, para ello expresamos dichas cantidades en la forma,

$$\begin{aligned}\overline{y}_M^{(j)} &= \frac{N-(n-l)}{N}\left[\overline{y}(m_j) - \frac{\overline{y}(m-m_j)}{\overline{x}(m-m_j)}\left(\overline{x}(m_j) - \frac{N\overline{X}-(n-l)\overline{x}(m-m_j)}{N-(n-l)}\right)\right] \\ &\quad + \frac{n-l}{N}\overline{y}(m-m_j)\end{aligned}$$

Por las propiedades del diseño muestral MAS(N,n), sabemos que que si $(n-l)$ unidades son seleccionadas al azar en m, las l unidades restantes forman una muestra aleatoria seleccionada mediante diseño MAS$(N-(n-l),l)$.

Así pues, descomponiendo la esperanza en dos fases, la primera fase, E_1, de obtención de l elementos de entre $N-(n-l)$, supuesto que $(n-l)$ unidades ya han sido seleccionadas; y la segunda fase, E_2, de selección de $(n-l)$ elementos a partir de N, obtenemos,

$$\begin{aligned}E[\overline{y}_M^{(j)}] &= E_2E_1[\overline{y}_M^{(j)}] \\ &= E_2\left[\frac{N-(n-l)}{N}\left(\frac{N\overline{Y}-(n-l)\overline{y}(m-m_j)}{N-(n-l)}\right) + \frac{n-l}{N}\overline{y}(m-m_j)\right] \\ &= E_2\left[\overline{Y}\right] = \overline{Y}\end{aligned}$$

■

4.4. Estrategias cuasi-insesgadas

Ya hemos visto que el estimador de razón, $\widehat{\overline{Y}}_R = (\overline{y}/\overline{x})\overline{X}$, en combinación con el diseño muestral aleatorio simple, da lugar a una estrategia sesgada, en el sentido de

que la estimación lo es, siendo el sesgo de orden $O(n^{-1})$. Denominaremos **estrategias cuasi-insesgadas** a aquellas que siendo sesgadas, proporcionan estimaciones con sesgo de orden $O(n^{-2})$ o inferior.

Las estrategias de este tipo, que vamos a considerar, están basadas en el diseño muestral aleatorio simple, MAS(N,n), en combinación con estimadores especiales, usualmente construidos a partir del estimador $\widehat{\overline{Y}}_R = (\overline{y}/\overline{x})\,\overline{X}$, de forma tal que el sesgo se reduzca en cierto orden de aproximación.

Así, el primer estimador que estudiamos se obtiene aplicando una técnica de tipo **jackknife**, al estimador $\widehat{\overline{Y}}_R = (\overline{y}/\overline{x})\,\overline{X}$. Para ello, utilizaremos las cantidades introducidas para construir el estimador de Mickey, y definiremos a partir de las mismas el siguiente estimador,

$$\widehat{\overline{Y}}_J = \frac{N-n+l}{N}\,k\widehat{\overline{Y}}_R - \frac{N-n}{N}\,\frac{k-1}{k}\sum_{j=1}^{k}\overline{y}_R^{(j)}$$

Observemos que si N es muy elevado, de forma que,

$$(N-n+l)/N \approx 1$$

$$(N-n)/N \approx 1$$

obtenemos la versión simplificada,

$$\widehat{\overline{Y}}'_J = k\widehat{\overline{Y}}_R - \frac{k-1}{k}\sum_{j=1}^{k}\overline{y}_R^{(j)}$$

que coincide con la forma de jackknife usual, tomando $k=n$, y por tanto $l=1$. Véase Quenouille (1956).

Con respecto a los anteriores estimadores, se tiene el siguiente resultado sobre su sesgo y error cuadrático medio, cuya demostración sigue la misma línea que la realizada para el estimador de Mickey.

Teorema 8 *El sesgo y el error cuadrático medio del estimador $\widehat{\overline{Y}}_J$ verifican,*

$$B[\widehat{\overline{Y}}_J] = O(n^{-2})$$

$$\mathrm{ECM}[\widehat{\overline{Y}}_J] = \frac{1-f}{n}\left(S_y^2 + \frac{\overline{Y}^2}{\overline{X}^2}S_x^2 - 2\frac{\overline{Y}}{\overline{X}}S_{xy}\right) + O(n^{-2})$$

y análogo para $\widehat{\overline{Y}}'_J$.

Este resultado nos dice que, en aproximación de primer orden, los estimadores $\widehat{\overline{Y}}_J$ e $\widehat{\overline{Y}}'_J$ son igualmente eficientes que $\widehat{\overline{Y}}_R = (\overline{y}/\overline{x})\,\overline{X}$. Véase Sukhatme *et al.* (1984).

Otras estrategias cuasi-insesgadas, basadas en diseño MAS(N,n), son las definidas por los siguientes estimadores,

- Estimador de De Pascual (1961),

$$\widehat{\overline{Y}}_{DP} = \widehat{\overline{Y}}_R + \frac{\overline{y} - \overline{zx}}{n-1} \qquad \text{con } Z_i = \frac{Y_i}{X_i}$$

- Estimador de Beale (1962),

$$\widehat{\overline{Y}}_B = \widehat{\overline{Y}}_R \left[1 + \left(\frac{1}{n} - \frac{1}{N}\right) \frac{s_{xy}}{\overline{xy}}\right] \left[1 + \left(\frac{1}{n} - \frac{1}{N}\right) \frac{s_x^2}{\overline{x^2}}\right]^{-1}$$

- Estimador de Tin (1965),

$$\widehat{\overline{Y}}_T = \widehat{\overline{Y}}_R \left[1 - \left(\frac{1}{n} - \frac{1}{N}\right) \left(\frac{s_x^2}{\overline{x^2}} - \frac{s_{xy}}{\overline{xy}}\right)\right]$$

Todos estos estimadores poseen un sesgo de orden $O(n^{-2})$, y con respecto a su eficiencia, se tiene el siguiente resultado,

Teorema 9 *Los errores cuadráticos medios de los estimadores de Tin, Beale y De Pascual verifican,*

$$\begin{aligned}
\mathrm{ECM}[\widehat{\overline{Y}}_T] &= \frac{1-f}{n}\left(S_y^2 + \frac{\overline{Y}^2}{\overline{X}^2}S_x^2 - 2\frac{\overline{Y}}{\overline{X}}S_{xy}\right) + O(n^{-2}) \\
\mathrm{ECM}[\widehat{\overline{Y}}_B] &= \frac{1-f}{n}\left(S_y^2 + \frac{\overline{Y}^2}{\overline{X}^2}S_x^2 - 2\frac{\overline{Y}}{\overline{X}}S_{xy}\right) + O(n^{-2}) \\
\mathrm{ECM}[\widehat{\overline{Y}}_{DP}] &= \frac{1-f}{n}\left(S_y^2 + \frac{\overline{Y}^2}{\overline{X}^2}S_x^2 - 2\frac{\overline{Y}}{\overline{X}}S_{xy}\right) + O(n^{-2})
\end{aligned}$$

Este teorema asegura que, en aproximación de primer orden, los estimadores de Tin, Beale y De Pascual son igualmente eficientes que el estimador de razón $\widehat{\overline{Y}}_R = (\overline{y}/\overline{x})\,\overline{X}$, y que el estimador $\widehat{\overline{Y}}_J$.

En las referencias citadas, y también en Williams (1961), Rao (1967), Rao y Beegle (1967) y Rao y Rao (1971) se pueden encontrar comparaciones, tanto analíticas como empíricas, entre estos estimadores. Algunas conclusiones que se desprenden de estos estudios comparativos son,

1. Bajo el modelo de proporcionalidad directa considerado en el enfoque predictivo, con función guía de la varianza $v(X_i) = X_i$, y tamaño muestral no muy reducido, es preferible el empleo del estimador clásico de razón, $\widehat{\overline{Y}}_R$.

2. El sesgo de $\widehat{\overline{Y}}_T$ es pequeño, y su error cuadrático medio es menor que para el resto de los estimadores, salvo para $\widehat{\overline{Y}}_R$ con $v(X_i) = X_i$.

3. El estimador de Beale, $\widehat{\overline{Y}}_B$, no difiere sustancialmente de $\widehat{\overline{Y}}_T$ salvo que n sea muy pequeño.

4. Si lo importante es la reducción del sesgo, y no necesariamente de error cuadrático medio, $\widehat{\overline{Y}}_J$ y $\widehat{\overline{Y}}_M$ han de ser preferidos al resto de los estimadores.

5. El estimador de Hartley-Ross, $\widehat{\overline{Y}}_{HR}$, puede no ser apropiado en ciertas circunstancias, bajo el modelo de proporcionalidad directa considerado. Hartley y Ross (1954), proponen para su estimador una modificación basada en la partición en grupos de la muestra.

Además de las referencias citadas, véanse también Sukhatme *et al.* (1984), Rao (1988) y Hedayat y Sinha (1991).

5. ESTIMADOR DE RAZÓN MULTIVARIANTE

El estimador de razón, $\widehat{\overline{Y}}_R = (\overline{y}/\overline{x})\,\overline{X}$, bajo diseño muestral aleatorio simple, se generaliza, de manera natural, al siguiente **estimador de razón multivariante**, propuesto por Olkin (1958),

$$\widehat{\overline{Y}}_{MR} = \overline{y}\sum_{i=1}^{p} w_i \frac{\overline{X}_i}{\overline{x}_i} = \sum_{i=1}^{p} w_i \widehat{\overline{Y}}_{Ri} \qquad \text{con } \widehat{\overline{Y}}_{Ri} = \frac{\overline{y}}{\overline{x}_i}\overline{X}_i$$

siendo $X_{i1}, X_{i2}, \ldots, X_{iN}$; $i = 1, \ldots, p$, p variables auxiliares conocidas, relacionadas con la variable de estudio, Y, con medias poblacionales respectivas $\overline{X}_i$, $i = 1, \ldots, p$, y medias muestrales $\overline{x}_i$, $i = 1, \ldots, p$. Las cantidades w_i son unos **pesos** que se determinarán en relación a la eficiencia del estimador.

Al ser combinación lineal de estimadores sesgados, $\widehat{\overline{Y}}_{MR}$ también lo será, y su sesgo estará influenciado por la elección de los pesos. En este sentido, se tiene el siguiente teorema, cuya demostración se basa en los resultados para el sesgo del estimador de razón univariante bajo diseño muestral MAS(N,n).

Teorema 10 *Una condición necesaria y suficiente para que el estimador $\widehat{\overline{Y}}_{MR}$ posea sesgo de orden $O(n^{-1})$ es que $\sum_{i=1}^{p} w_i = 1$, siendo en tal caso,*

$$B[\widehat{\overline{Y}}_{MR}] = \frac{1-f}{n}\sum_{i=1}^{p} w_i \left(\frac{\overline{Y}}{\overline{X}_i^2} S_{xi}^2 - \frac{1}{\overline{X}_i} S_{yxi} \right) + O(n^{-2})$$

donde S_{xi}^2 denota la cuasivarianza poblacional de $X_{i1},\ldots,X_{iN}$, y S_{yxi} la correspondiente cuasicovarianza.

En lo que sigue, supondremos ya que $\sum_{i=1}^{p} w_i = 1$, con objeto de controlar el sesgo.

Para el estudio del error cuadrático medio utilizaremos la siguiente expresión, que se obtiene mediante un cálculo directo,

$$E\left[(\widehat{\overline{Y}}_{Ri} - \overline{Y})(\widehat{\overline{Y}}_{Rj} - \overline{Y})\right] = \frac{1-f}{n}\left(S_y^2 - \frac{\overline{Y}}{\overline{X}_i} S_{yxi} - \frac{\overline{Y}}{\overline{X}_j} S_{yxj} + \frac{\overline{Y}^2}{\overline{X}_i \overline{X}_j} S_{xixj} \right) + O(n^{-2})$$

verificándose el siguiente resultado,

Teorema 11 *El error cuadrático medio del estimador multivariante de razón, $\widehat{\overline{Y}}_{MR}$, es de orden $O(n^{-1})$, siendo además,*

$$\mathrm{ECM}[\widehat{\overline{Y}}_{MR}] = \frac{1-f}{n}\sum_{i,j=1}^{p} w_i w_j \left(S_y^2 - \frac{\overline{Y}}{\overline{X}_i} S_{yxi} - \frac{\overline{Y}}{\overline{X}_j} S_{yxj} + \frac{\overline{Y}^2}{\overline{X}_i \overline{X}_j} S_{xixj} \right) + O(n^{-2})$$

Demostración: Al ser $\sum_{i=1}^{p} w_i = 1$, podemos escribir,

$$\begin{aligned}
\mathrm{ECM}[\widehat{\overline{Y}}_{MR}] &= E\left[\left(\sum_{i=1}^{p} w_i \widehat{\overline{Y}}_{Ri} - \overline{Y} \right)^2 \right] = E\left[\left(\sum_{i=1}^{p} w_i \left(\widehat{\overline{Y}}_{Ri} - \overline{Y} \right) \right)^2 \right] \\
&= E\left[\sum_{i=1}^{p} w_i^2 \left(\widehat{\overline{Y}}_{Ri} - \overline{Y} \right)^2 + \sum_{i \neq j=1}^{p}\sum w_i w_j \left(\widehat{\overline{Y}}_{Ri} - \overline{Y} \right)\left(\widehat{\overline{Y}}_{Rj} - \overline{Y} \right) \right] \\
&= \sum_{i=1}^{p} w_i^2 \,\mathrm{ECM}\left[\widehat{\overline{Y}}_{Ri} \right] + \sum_{i \neq j=1}^{p}\sum w_i w_j E\left[\left(\widehat{\overline{Y}}_{Ri} - \overline{Y} \right)\left(\widehat{\overline{Y}}_{Rj} - \overline{Y} \right) \right]
\end{aligned}$$

Y basta aplicar los resultados para el error cuadrático medio del estimador de razón univariante, así como la expresión de $E\left[(\widehat{\overline{Y}}_{Ri}-\overline{Y})(\widehat{\overline{Y}}_{Rj}-\overline{Y})\right]$, para obtener el resultado propuesto.

■

Planteamos ahora el problema de calcular los pesos de forma que el error sea lo menor posible. Para ello trabajaremos con la aproximación de primer orden del error cuadrático medio, es decir,

$$\mathrm{ECM}_1[\widehat{\overline{Y}}_{MR}] = \frac{1-f}{n}\sum_{i,j=1}^{p} w_i w_j \left(S_y^2 - \frac{\overline{Y}}{\overline{X}_i}S_{yxi} - \frac{\overline{Y}}{\overline{X}_j}S_{yxj} + \frac{\overline{Y}^2}{\overline{X}_i\overline{X}_j}S_{xixj}\right)$$

que se puede expresar como $\mathrm{ECM}_1[\widehat{\overline{Y}}_{MR}] = wAw'$ siendo $w = (w_1,\ldots,w_p)$, y $A = (a_{ij})$ la matriz de dimensión $p \times p$ definida como,

$$a_{ij} = \frac{1-f}{n}\left(S_y^2 - \frac{\overline{Y}}{\overline{X}_i}S_{yxi} - \frac{\overline{Y}}{\overline{X}_j}S_{yxj} + \frac{\overline{Y}^2}{\overline{X}_i\overline{X}_j}S_{xixj}\right) \qquad i,j = 1,\ldots,p$$

Observemos que A es simétrica y definida positiva, luego, por la desigualdad de Cauchy generalizada, se tiene,

$$(ab')^2 \leq (aAa')(bA^{-1}b') \qquad \forall a,b \in R^p$$

donde la igualdad se da si y sólo si $aA = \alpha b$ siendo α un escalar no nulo. En particular, tomando $a = w$, y $b = e = (1,\ldots,1)$, tendremos,

$$1 = (we')^2 \leq (wAw')(eA^{-1}e')$$

con lo que (wAw') toma el valor mínimo si y sólo si $wA = \alpha e$, es decir, $w = \alpha eA^{-1}$, y a partir de la condición de normalidad de los pesos, se obtiene el siguiente vector de pesos **óptimos**,

$$w_{\mathrm{opt}} = \frac{eA^{-1}}{eA^{-1}e'}$$

y con esta elección, el término guía, es decir, de orden $O(n^{-1})$ del error cuadrático medio resulta ser,

$$\mathrm{ECM}_1[\widehat{\overline{Y}}_{MR}] = \frac{1}{eA^{-1}e'}$$

Además de la fundamental referencia de Olkin, pueden consultarse Sukhatme *et al.* (1984), donde se realiza un tratamiento exhaustivo del caso $p = 2$, y Cochran (1993).

Hemos de observar también que si la correlación entre la variable Y y las variables auxiliares es positiva para algunas y negativas para otras, se puede plantear un estimador multivariante mixto basado en una combinación lineal de estimadores de razón y producto, propuesto por Rao y Mudholkar (1967), tomando como base las propiedades del estimador de producto de Murthy (1964).

6. ESTIMADOR DE RAZÓN Y ESTRATIFICACIÓN

En este apartado estudiaremos las formas que adopta el estimador usual de razón cuando se utiliza muestreo estratificado, y más concretamente, diseño muestral aleatorio simple estratificado, MASE(N,n), consistente en realizar diferentes muestreos aleatorios simples en cada uno de los estratos, siendo $N = (N_1,\ldots,N_L)$ un vector cuyas componentes son los tamaños de los diferentes estratos, y lo mismo para $n = (n_1,\ldots,n_L)$, compuesto por los tamaños de muestra en cada estrato.

Dependiendo de la metodología aplicada, obtendremos dos tipos de estimación, **separada** y **combinada**.

6.1. Estimación separada de razón

Teniendo en cuenta la descomposición usual de la media poblacional como combinación lineal de medias en cada estrato,

$$\overline{Y} = \sum_{h=1}^{L} W_h \overline{Y}_h$$

podemos estimar las medias de cada estrato empleando estimadores de razón, obteniendo el estimador separado,

$$\widehat{\overline{Y}}_{RS} = \sum_{h=1}^{L} W_h \widehat{\overline{Y}}_{hR}$$

Obviamente, las propiedades de este estimador dependerán de los estimadores de razón utilizados en los distintos estratos. En particular podrá ser sesgado, insesgado o cuasi-insesgado, y su eficiencia, medida por ejemplo en términos del error cuadrático medio, estará influenciada por la eficiencia de las estimaciones en cada estrato, ya que, por la independencia de las extracciones, se tendrá,

$$\mathrm{ECM}[\widehat{\overline{Y}}_{RS}] = \sum_{h=1}^{L} W_h^2 \,\mathrm{ECM}\left[\widehat{\overline{Y}}_{hR}\right]$$

En particular, si empleamos los estimadores de razón usuales para el muestreo aleatorio,

$$\widehat{\overline{Y}}_{hR} = (\overline{y}_h/\overline{x}_h)\overline{X}_h = \widehat{R}_h\overline{X}_k \qquad h = 1,\ldots,L$$

obtendremos,

$$\widehat{\overline{Y}}_{RS} = \sum_{h=1}^{L} W_h \frac{\overline{y}_h}{\overline{x}_h}\overline{X}_h$$

que será sesgado, siendo su sesgo,

$$\begin{aligned} B[\widehat{\overline{Y}}_{RS}] &= E\left[\sum_{h=1}^{L} W_h \frac{\overline{y}_h}{\overline{x}_h}\overline{X}_h\right] - \sum_{h=1}^{L} W_h\overline{Y}_h \\ &= E\left[\sum_{h=1}^{L} W_h\left(\frac{\overline{y}_h}{\overline{x}_h}\overline{X}_h - \overline{Y}_h\right)\right] = \sum_{h=1}^{L} W_h\left(E[\widehat{R}_h]E[\overline{x}_h] - E[\widehat{R}_h\overline{x}_h]\right) \\ &= -\sum_{h=1}^{L} W_h \operatorname{Cov}[\widehat{R}_h,\overline{x}_h] \end{aligned}$$

y suponiendo para los coeficientes de variación de $\overline{x}_h$ la acotación común $\mathrm{CV}[\overline{x}_h] = \sigma[\overline{x}_h]/E[\overline{x}_h] \leq C_0,\ \forall h$, obtenemos,

$$\left|B[\widehat{\overline{Y}}_{RS}]\right| \leq \sum_{h=1}^{L} W_h\sigma[\widehat{\overline{Y}}_{hR}]\mathrm{CV}[\overline{x}_h] \leq C_0\sum_{h=1}^{L} W_h\sigma[\widehat{\overline{Y}}_{hR}]$$

que proporciona la siguiente cota,

$$\frac{\left|B[\widehat{\overline{Y}}_{RS}]\right|}{\sigma[\widehat{\overline{Y}}_{RS}]} \leq C_0\sqrt{L}$$

Podemos pues afirmar que si el número de estratos es elevado, el sesgo de la estimación puede llegar a ser apreciable. Ello sugiere el empleo de estimadores insesgados o cuasi-insesgados, sobre todo si los tamaños de la muestra en los diferentes estratos no son muy elevados.

Con respecto al error cuadrático medio de la estimación anterior, tendremos, en primer orden de aproximación,

$$\mathrm{ECM}_1[\widehat{\overline{Y}}_{RS}] = \sum_{h=1}^{L} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right)\left(S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{xyh}\right)$$

6.2. Estimación combinada de razón

Otra posibilidad para reducir el sesgo es emplear el estimador combinado de razón, propuesto por Hansen, Hurwitz y Gurney (1946), y definido como,

$$\widehat{\overline{Y}}_{RC} = \frac{\sum_{h=1}^{L} W_h \overline{y}_h}{\sum_{h=1}^{L} W_h \overline{x}_h}\overline{X} = \frac{\overline{y}_{\mathrm{est}}}{\overline{x}_{\mathrm{est}}}\overline{X} = \widehat{R}_{\mathrm{est}}\overline{X}$$

Este estimador también es sesgado, siendo ahora el sesgo,

$$\begin{aligned} B[\widehat{\overline{Y}}_{RC}] &= E[\widehat{R}_{\mathrm{est}}]\overline{X} - \overline{Y} \\ &= E[\widehat{R}_{\mathrm{est}}]E[\overline{x}_{\mathrm{est}}] - E[\widehat{R}_{\mathrm{est}}\overline{x}_{\mathrm{est}}] \\ &= -\mathrm{Cov}[\widehat{R}_{\mathrm{est}}, \overline{x}_{\mathrm{est}}] \end{aligned}$$

de donde se obtiene,

$$\frac{\left|B[\widehat{\overline{Y}}_{RC}]\right|}{\sigma[\widehat{\overline{Y}}_{RC}]} \leq \mathrm{CV}[\overline{x}_{\mathrm{est}}]$$

es decir, si el coeficiente de variación de $\overline{x}_{\mathrm{est}}$ es pequeño, el sesgo puede ser despreciable.

Para el error cuadrático medio del estimador combinado tendremos, en primer orden de aproximación,

$$\mathrm{ECM}_1[\widehat{\overline{Y}}_{RC}] = \sum_{h=1}^{L} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right)\left(S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{xyh}\right)$$

Es interesante comparar este error cuadrático medio con el correspondiente del estimador separado. De esta forma, es inmediato obtener,

$$\begin{aligned} \mathrm{ECM}_1[\widehat{\overline{Y}}_{RC}] - \mathrm{ECM}_1[\widehat{\overline{Y}}_{RS}] &= \sum_{h=1}^{L} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right)\left(S_{xh}^2 (R - R_h)^2 \right. \\ &\quad \left. + 2(R - R_h)(R_h S_{xh}^2 - S_{xyh})\right) \end{aligned}$$

Como puede verse, la diferencia depende de la variabilidad de las razones, R_h, en los diferentes estratos, y también de las cantidades $(R_h S_{xh}^2 - S_{xyh})$. Estas últimas serán, usualmente, pequeñas, pudiendo ser despreciables si el ajuste al modelo de proporcionalidad directa es muy satisfactorio en todos los estratos. Así pues, el estimador separado resulta ser más eficiente que el combinado, a menos que las razones R_h presenten muy poca variabilidad a lo largo de los estratos. Véanse Sukhatme *et al.* (1984), Cochran (1993) y para ejemplos numéricos, Fernández y Mayor (1995).

7. OPTIMALIDAD DEL ESTIMADOR DE RAZÓN

Ya hemos visto como, en general, los estimadores de razón no son insesgados **en el diseño**, esto es, no tienen por qué cumplir,

$$E_d\left[\widehat{\overline{Y}}_R\right] = \sum_{m \in M} p(m)\widehat{\overline{Y}}_R(m) = \overline{Y}$$

lo que, en principio, los descarta como candidatos a ser considerados óptimos. No obstante, sí es posible considerar, bajo un determinado modelo de superpoblación, la denominada insesgadez **en el modelo**, según la definición que exponemos a continuación.

Definición *Dado un diseño muestral, $d = (M, p(\cdot))$, y un modelo de superpoblación, S, diremos que el estimador de $\theta(Y)$, $\widehat{\theta}$, es insesgado en el modelo S, si cumple,*

$$E_s\left[\widehat{\theta} - \theta(Y)\right] = 0 \qquad \forall m \in M$$

Por ejemplo, si consideramos el modelo de proporcionalidad directa ya empleado para obtener el estimador de razón genérico,

$$\begin{aligned} &Y_i = \beta X_i + \varepsilon_i \\ &\\ &E_s[\varepsilon_i] = 0 \\ &V_s[\varepsilon_i] = \sigma^2 v(X_i) \\ &E_s[\varepsilon_i \varepsilon_j] = 0, \ i \neq j \end{aligned}$$

se tendrá, con $v(x) = x$,

$$E_s\left[\widehat{\overline{Y}}_R - \overline{Y}\right] = E_s\left[\frac{\sum_{i\in m} Y_i/\pi_i}{\sum_{i\in m} X_i/\pi_i}\overline{X} - \overline{Y}\right] = \frac{\sum_{i\in m} \beta X_i/\pi_i}{\sum_{i\in m} X_i/\pi_i}\overline{X} - \beta\overline{X} = 0$$

es decir, el estimador de razón genérico es insesgado en el modelo, aunque , en general, no lo sea en el diseño. Es obvio que este nuevo concepto de insesgadez se puede considerar más débil que el usual, en el sentido de que está supeditado a que el modelo de superpoblación sea el adecuado. Observemos también que, al igual que un estimador puede ser insesgado en el modelo pero no en el diseño, es posible que sea insesgado en el diseño pero no en el modelo. Por ejemplo, en el diseño $\mathrm{MAS}(N,n)$, la media muestral, $\bar{y}$ es un estimador insesgado (en el diseño) de $\overline{Y}$, sin embargo, para el modelo de proporcionalidad directa se tiene,

$$E_s[\bar{y} - \overline{Y}] = \beta\bar{x} - \beta\overline{X} = \beta(\bar{x} - \overline{X})$$

que, en general, es distinto de cero.

Bajo este enfoque, que podemos denominar **dependiente del modelo** es posible obtener ciertas propiedades de optimalidad sobre el estimador de razón, que estudiaremos a continuación, siguiendo la línea iniciada en Royall (1970).

Para simplificar la notación, supondremos que el parámetro a estimar es el total poblacional, $T(Y) = \sum_{i\in U} Y_i$, y que el estimador, $\widehat{T}(Y)$, se va a buscar en la clase de los estimadores lineales e insesgados en el modelo.

Observemos que, dada una muestra, m, perteneciente al diseño muestral que estemos empleando, $\widehat{T}(Y)$ admite la descomposición,

$$\widehat{T}(Y) = \sum_{i\in m} Y_i + \frac{\widehat{T}(Y) - \sum_{i\in m} Y_i}{\sum_{i\in U-m} X_i} \sum_{i\in U-m} X_i = \sum_{i\in m} Y_i + \widehat{\beta} \sum_{i\in U-m} X_i$$

donde la notación introducida se justifica al ser $E_s[\widehat{\beta}] = \beta$, lo que se comprueba fácilmente. Además, al ser $\widehat{T}(Y)$ lineal en $Y_i,\ i \in m$, también lo será $\widehat{\beta}$, es decir,

$$\widehat{\beta} = \sum_{i\in m} \alpha_i Y_i$$

Dado un diseño muestral, $d = (M, p(\cdot))$, podemos utilizar la siguiente cantidad,

$$\mathrm{ECM}[\widehat{T}(Y), d] = E_s \mathrm{ECM}_d[\widehat{T}(Y)] = E_s\left[\sum_{m\in M}\left(\widehat{T}(m,Y) - T(Y)\right)^2 p(m)\right]$$

como medida de la eficiencia del estimador $\widehat{T}(Y)$. Esta medida involucra tanto el efecto del diseño muestral, como la influencia del modelo de superpoblación, y en relación a la misma, se tiene el siguiente resultado,

Lema *Para un diseño muestral cualquiera, $d = (M, p(\cdot))$, y los estimadores del total,*

$$\begin{aligned}\widehat{T}'(Y) &= \sum_{i\in m} Y_i + \widehat{\beta}' \sum_{i\in U-m} X_i \\ \widehat{T}''(Y) &= \sum_{i\in m} Y_i + \widehat{\beta}'' \sum_{i\in U-m} X_i\end{aligned}$$

se verifica que si,

$$E_s\left[\left(\widehat{\beta}' - \beta\right)^2\right] \leq E_s\left[\left(\widehat{\beta}'' - \beta\right)^2\right] \qquad \forall m \in M$$

entonces,

$$\mathrm{ECM}\left[\widehat{T}'(Y), d\right] \leq \mathrm{ECM}\left[\widehat{T}''(Y), d\right]$$

Demostración: Teniendo en cuenta el modelo de superpoblación, podemos realizar el siguiente desarrollo,

$$\begin{aligned}\mathrm{ECM}\left[\widehat{T}'(Y), d\right] &= E_s\left[\sum_{m\in M}\left(\widehat{T}(m,Y) - T(Y)\right)^2 p(m)\right] \\ &= \sum_{m\in M} E_s\left[\left(\widehat{T}(m,Y) - T(Y)\right)^2\right] p(m) \\ &= \sum_{m\in M} E_s\left[\left(\sum_{i\in m} Y_i + \widehat{\beta}' \sum_{i\in U-m} X_i - \sum_{i\in m} Y_i - \sum_{i\in U-m} Y_i\right)^2\right] p(m) \\ &= \sum_{m\in M} E_s\left[\left(\widehat{\beta}' \sum_{i\in U-m} X_i - \beta \sum_{i\in U-m} X_i - \sum_{i\in U-m} \varepsilon_i\right)^2\right] p(m) \\ &= \sum_{m\in M}\left(\left(\sum_{i\in U-m} X_i\right)^2 E_s\left[\left(\widehat{\beta}' - \beta\right)^2\right] + \sigma^2 \sum_{i\in U-m} v(X_i)\right) p(m)\end{aligned}$$

de donde se deduce inmediatamente el resultado propuesto.

■

Con respecto a la optimalidad del estimador de razón, se verifica el siguiente resultado fundamental, en cuya demostración emplearemos el anterior lema.

Teorema 12 *Dado un diseño muestral cualquiera, $d = (M, p(\cdot))$, sea,*

$$\widehat{\beta}^* = \frac{\sum_{i\in m} X_i Y_i / v(X_i)}{\sum_{i\in m} X_i^2 / v(X_i)}$$

y el estimador del total,

$$\widehat{T}^*(Y) = \sum_{i\in m} Y_i + \widehat{\beta}^* \sum_{i\in U-m} X_i$$

Entonces se verifica que para cualquier estimador del total, $\widehat{T}(Y)$, lineal e insesgado en el modelo,

$$\mathrm{ECM}[\widehat{T}^*(Y), d] \leq \mathrm{ECM}[\widehat{T}(Y), d]$$

siendo además,

$$\mathrm{ECM}[\widehat{T}^*(Y), d] = \sigma^2 \sum_{m\in M} \left(\frac{\left(\sum_{i\in U-m} X_i\right)^2}{\sum_{i\in m} X_i^2 / v(X_i)} + \sum_{i\in U-m} v(X_i) \right) p(m)$$

Demostración: Para la demostración, nos basaremos en el lema anterior. Sea pues m una muestra del diseño, y $\widehat{\beta}$ lineal en $Y_i,\ i \in m$, esto es,

$$\widehat{\beta} = \sum_{i\in m} \alpha_i Y_i$$

A partir de la condición $E_s[\widehat{\beta}] = \beta$ se obtiene la restricción,

$$\sum_{i\in m} \alpha_i X_i = 1$$

Por otra parte,

$$\begin{aligned} E_s\left[\left(\widehat{\beta} - \beta\right)^2\right] &= E_s\left[\left(\sum_{i\in m} \alpha_i (Y_i - \beta X_i)\right)^2\right] \\ &= E_s\left[\sum_{i,j\in m} \alpha_i \alpha_j (Y_i - \beta X_i)(Y_j - \beta X_j)\right] \\ &= E_s\left[\sum_{i\in m} \alpha_i^2 (Y_i - \beta X_i)^2\right] = \sigma^2 \sum_{i\in m} \alpha_i^2 v(X_i) \end{aligned}$$

que, con la restricción ya obtenida anteriormente,

$$\sum_{i\in m} \alpha_i X_i = 1$$

alcanza el mínimo para los valores,

$$\alpha_i = \frac{X_i/v(X_i)}{\sum_{i\in m} X_i^2/v(X_i)} \qquad i \in m$$

lo que proporciona precisamente $\widehat{\beta}^*$. Hemos demostrado pues que $\forall m \in M$,

$$E_s\left[\left(\widehat{\beta}^* - \beta\right)^2\right] \leq E_s\left[\left(\widehat{\beta} - \beta\right)^2\right] \qquad \forall \widehat{\beta}$$

y basta tener en cuenta el anterior lema para obtener el resultado de optimalidad enunciado.

Finalmente, la expresión para $\mathrm{ECM}[\widehat{T}^*(Y), d]$ se obtiene mediante un cálculo directo.

■

Observemos que para el caso $v(x) = x$, ya considerado en el enfoque predictivo aplicado al principio del tema, se obtiene como estimador óptimo,

$$\widehat{T}^*(Y) = \sum_{i\in m} Y_i + \frac{\sum_{i\in m} Y_i X_i/X_i}{\sum_{i\in m} X_i^2/X_i} \sum_{i\in U-m} X_i = \frac{\overline{y}}{\overline{x}} T(X)$$

Este resultado no está en contradicción con las buenas propiedades que presenta el estimador genérico de razón obtenido en la sección 3. de esta revisión,

$$\widehat{T}_R(Y) = \frac{\sum_{i\in m} Y_i/\pi_i}{\sum_{i\in m} X_i/\pi_i} T(X)$$

en lo que respecta a la estimación del error. Por otra parte, hay que tener en cuenta que el estimador óptimo ha sido buscado en una clase muy especial de estimadores lineales, **insesgados para el modelo**, con los inconvenientes de dependencia del mismo que ello supone.

Observemos también que para el caso $v(x) = x$, el error cuadrático medio resulta ser,

$$\mathrm{ECM}[\widehat{T}^*(Y), d] \quad = \sigma^2 \sum_{m\in M} \left(\frac{\left(\sum_{i\in U-m} X_i\right)^2}{\sum_{i\in m} X_i^2/X_i} + \sum_{i\in U-m} X_i \right) p(m)$$

$$= \sigma^2 T(X) E_d\left[\frac{\sum_{i\in U-m} X_i}{\sum_{i\in m} X_i}\right]$$

Y si denotamos por m^* la muestra formada por los elementos de la población, tal que,

$$\sum_{i \in m^*} X_i = \max_{m \in M} \left\{ \sum_{i \in m} X_i \right\}$$

el anterior error cuadrático medio será mínimo para el siguiente diseño muestral **intencional**,

$$d^* = (\{m^*\}, p(m^*) = 1)$$

es decir, un diseño con una única muestra, m^*, que es seleccionada con probabilidad uno. Hemos obtenido pues el siguiente resultado,

Teorema 13 *Bajo el modelo de superpoblación de proporcionalidad directa, con función guía de la varianza $v(x) = x$, la estrategia muestral $(d^*, (\overline{y}/\overline{x})\, T(X))$ es óptima para la estimación de $T(Y)$.*

Notemos que el resultado anterior sigue siendo válido si solamente exigimos que $v(x)$ sea no decreciente, y $v(x)/x^2$ no creciente.

A pesar de su importancia teórica, estos resultados, como afirman Cassel, Särndal y Wretman (1977), están en conflicto con uno de los principios más extendidos de la Estadística como es el de la aleatorización. Por otra parte, la estrategia óptima anterior es incompatible con el cálculo o la estimación del error.

Como fuentes importantes para profundizar en estas cuestiones, citaremos Royall (1970), Royall y Herson (1973a, 1973b), Royall y Eberhardt (1975), Cassel, Särndal y Wretman (1977), Bellhouse (1984), Sukhatme *et al.* (1984) y Chaudhuri y Vos (1988).

Observemos finalmente que es posible plantear el problema complementario de hallar las probabilidades de inclusión óptimas para que el estimador de razón genérico,

$$\widehat{T}_R(Y) = \frac{\sum_{i \in m} Y_i/\pi_i}{\sum_{i \in m} X_i/\pi_i} T(X)$$

sea óptimo bajo el modelo de superpoblación de proporcionalidad directa. Véase Särndal, Swensson y Wretman (1992) para un estudio del mismo.

8. REFERENCIAS

[1] **Azorín, F.** y **Sánchez-Crespo, J.L.** (1986). *Métodos y Aplicaciones del Muestreo*. Alianza Universidad Textos. Madrid.

[2] **Beale, E.M.L.** (1962). «Some use of computers in operational research». *Industrielle Organization,* **31**, 27–28.

[3] **Bellhouse, D.R.** (1984). «A review of optimal designs in survey sampling». *The Canadian Journal of Statistics,* **12**, 53–65.

[4] **Cassel, C., Särndal, C.** y **Wretman, J.** (1977). *Foundations of Inference in Survey Sampling.* Wiley. New York.

[5] **Cochran, W.G.** (1978). «Laplace's Ratio Estimator». *Contributions to Survey Sampling and Applied Statistics.* H.A. David (ed.). Academic Press. New York.

[6] **Cochran, W.G.** (1993). *Técnicas de Muestreo.* Décima reimpresión. CECSA. México.

[7] **Chang, W.C.** (1976). «Statistical theories and sampling practice». *On the History of Statistics and Probability.* D.B. Owen (ed.). Dekker. New York.

[8] **Chaudhuri, A.** y **Vos, J.** (1988). *Unified Theory and Strategies of Survey Sampling.* North Holland. Amsterdam.

[9] **David, I.P.** y **Sukhatme, B.V.** (1974). «On the bias and mean square error of the ratio estimator». *J. Amer. Statist. Assoc.*, **69**, 464–466.

[10] **De Pascual, N.** (1961). «Unbiased ratio estimators in stratified sampling». *J. Amer. Statist. Assoc.,* **56**, 70–87.

[11] **Fernández, F.R.** y **Mayor, J.A.** (1995). *Muestreo en Poblaciones Finitas: Curso Básico*. E.U.B. Barcelona. (También en P.P.U., (1994)).

[12] **Fuller, W.A.** (1975). «Regression analysis for sample survey». *Sankhyā,* **C37**, 117–132.

[13] **Goodman, L.A.** y **Hartley, H.O.** (1958). «The precision of unbiased ratio-type estimators». *J. Amer. Statist. Assoc.,* **53**, 491–508.

[14] **Hald, A.** (1990). *A History of Probability and Statistics and Their Applications before 1750.* Wiley. New York.

[15] **Hansen, M.H., Hurwitz, W.N.** y **Gurney, M.** (1946). «Problems and methods of the sample survey of business». *J. Amer. Statist. Assoc.*, **41**, 173–189.

[16] **Hansen, M.H., Hurwitz, W.N.** y **Madow, W.G.** (1953). *Sample Survey Methods and Theory. Vol. I y II.* Wiley. New York.

[17] **Hartley, H.O.** y **Ross, A.** (1954). «Unbiased ratio estimators». *Nature,* **174**, 270–271.

[18] **Hedayat, A.S.** y **Sinha, B.K.** (1991). *Design and Inference in Finite Population Sampling.* Wiley. New York.

[19] **Kish, L.** y **Frankel, M.R.** (1974). «Inference from complex samples». *J. Roy. Statist. Soc.,* **B36**, 1–37.

[20] **Lahiri, D.B.** (1951). «A method of sample selection providing unbiased ratio estimates». *Bulletin of the International Statistical Institute*, **33**, 133–140.

[21] **Lanke, J.** (1974). «On nonnegative variance estimators in survey sampling». *Sankhyā*, **C36**, 33–42.

[22] **Mickey, M.R.** (1959). «Some finite population unbiased ratio and regression estimators». *J. Amer. Statist. Assoc.*, **54**, 594–612.

[23] **Midzuno, H.** (1952). «On the sampling system with probability proportionate to sum of sizes». *Annals of the Institute of Statistical Mathematics*, **3**, 99–107.

[24] **Murthy, M.N.** (1964). «Product method of estimation». *Sankhyā*, **A26**, 69–74.

[25] **Olkin, I.** (1958). «Multivariate ratio estimation for finite populations». *Biometrika*, **45**, 154–165.

[26] **Quenouille, M.H.** (1956). «Notes on bias in estimation». *Biometrika*, **43**, 353–360.

[27] **Rao, C.R.** (1965). *Linear Statistical Inference and its Applications.* Wiley. New York.

[28] **Rao, J.N.K.** (1967). «The precision of Mickey's unbiased ratio estimator». *Biometrika*, **54**, 321–324.

[29] **Rao, J.N.K.** y **Beegle, L.D.** (1977). «A Monte Carlo study of some ratio estimators». *Sankhyā*, **B29**, 47–56.

[30] **Rao, J.N.K.** y **Vijayan, K.** (1977). «On estimating the variance in sampling with probability proportional to aggregate size». *J. Amer. Statist. Assoc.*, **72**, 579–584.

[31] **Rao, P.S.R.S.** (1971). «Small sample results for ratio estimators». *Biometrika*, **58**, 625–630.

[32] **Rao, P.S.R.S.** (1988). «Ratio and regression estimators». *Handbook of Statistics 6. Sampling.* Krishnaiah y Rao, (Eds.). North Holland. Amsterdam.

[33] **Rao, P.S.R.S.** y **Mudholkar, G.S.** (1967). «Generalized multivariate estimator for the mean of finite populations». *J. Amer. Statist. Assoc.*, **62**, 1009–1012.

[34] **Rao, T.J.** (1966). «On the variance of the ratio estimator for Midzuno-Sen sampling scheme». *Metrika*, **10**, 89–91.

[35] **Rao, T.J.** (1972). «On the variance of the ratio estimator». *Metrika*, **18**, 209–215.

[36] **Robson, D.S.** (1957). «Application of multivariate polykays to the theory of unbiased ratio-type estimators». *J. Amer. Statist. Assoc.*, **52**, 511–522.

[37] **Royall, R.M.** (1970). «On finite population sampling theory under certain linear regression models». *Biometrika*, **57**, 377–387.

[38] **Royall, R.M.** y **Herson, J.** (1973a). «Robust estimation in finite populations I». *J. Amer. Statist. Assoc.*, **68**, 880–890.

[39] **Royall, R.M.** y **Herson, J.** (1973b). «Robust estimation in finite populations II». *J. Amer. Statist. Assoc.*, **68**, 890–894.

[40] **Royall, R.M.** y **Eberhardt, K.R.** (1975). «Variance estimates for the ratio estimator». *Sankhyā*, **C37**, 43–52.

[41] **Royall, R.M.** y **Cumberland, W.G.** (1981). «An empirical study of the ratio estimator and estimators of its variance». *J. Amer. Statist. Assoc.*, **76**, 66–77.

[42] **Ruiz, M.** y **Santos, J.** (1989). «Unbiased mean-of-the-ratios estimators». *Statistica*, **49**, 617–622.

[43] **Sahoo, L.N.** y **Ruiz, M.** (1994). «Unbiased estimators using auxiliary information in sample surveys: a review». *Revista de la Academia de Ciencias de Zaragoza*, **49**, 137–146.

[44] **Sánchez-Crespo, J.L.** (1980). *Curso Intensivo de Muestreo en Poblaciones Finitas.* 2^a edición. Instituto Nacional de Estadística. Madrid.

[45] **Särndal, C.** (1984). *Inférence Statistique et Analyse des Données sous des Plans d'Échantillonnage Complexes.* Presses de l'Université de Montréal.

[46] **Särndal, C., Swensson, B.** y **Wretman, J.** (1992). *Model Assisted Survey Sampling.* Springer-Verlag. New York.

[47] **Singh, P.** y **Srivastava, A.K.** (1980). «Sampling schemes providing unbiased regression estimators». *Biometrika*, **67**, 205–209.

[48] **Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S.** y **Asok, C.** (1984). *Sampling Theory of Surveys Applications.* Tercera edición. Iowa State University Press. Ames. Iowa.

[49] **Tin, M.** (1965). «Comparison of some ratio estimators». *J. Amer. Statist. Assoc.*, **60**, 294–307.

[50] **Williams, W.H.** (1961). «Generating unbiased ratio and regression estimators». *Biometrics*, **17**, 267–274.

ENGLISH SUMMARY

RATIO ESTIMATORS: A REVIEW

JOSÉ A. MAYOR GALLEGO*
Universidad de Sevilla

In this review we expose the basic principles relating to the use of auxiliary information, in order to estimate linear parameters defined over finite populations, by means of ratio type estimators. The construction of these estimators is firstly carried out by means of an heuristic approach based on the existence of a direct proportionality relation between the study variable and the auxiliary variable, but a more formal study is carried out under a superpulation model approach. The main problem of these estimators is the existence of bias, and in order to reduce it, we have to use special sampling designs or to modify the structure of the estimators to obtain unbiased or almost unbiased strategies.

Keywords: Sampling, finite populations, ratio-type estimators

AMS Classification: 62D05

*José A. Mayor Gallego. Dpto. Estadística e Investigación Operativa. Universidad de Sevilla. C/Tarfia s/n. 41012 SEVILLA.

– Received may 1996.

– Accepted april 1997.

A way to integrate the auxiliary information, in order to increase the accuracy of the estimates over a finite population is to use estimators with a rational mathematical structure, involving the study variable, Y, and an auxiliary variable, X, completely known.

In this paper, we **review** this class of estimators, emphasizing the importance of the proportionality between the X and Y variables, in order to obtain good estimates, but controlling the bias.

We are going to suppose that we are estimating the linear parameter,

$$\theta(Y) = \sum_{i \in U} a_i Y_i$$

over a finite population, $U = \{1, 2, \ldots, N\}$, by means of a sample, s, chosen from a sampling design, d.

Thus we study firstly the **heuristic approach**, based on the following factorizing, (for the population mean),

$$\overline{Y} = \frac{\overline{Y}}{\overline{X}} \overline{X} = R\overline{X}$$

then, we can define the estimator,

$$\widehat{\overline{Y}}_R = \widehat{R}\overline{X}$$

The final form of this estimator depends on the sampling design. If we use the simple random sampling, $\mathrm{SRS}(N, n)$, we can estimate R using the sample means ratio, that is to say,

$$\widehat{\overline{Y}}_R = \frac{\overline{y}}{\overline{x}} \overline{X}$$

The efficiency of this estimator depends on the relation between the study variable, Y, and the auxiliary variable, X. It has a good behaviour if there is an approximate relation, $Y \approx \alpha X$, that is to say, a direct proportionality. Using approximate techniques, we obtain the following expressions for the variance and its estimation,

$$\begin{aligned} V[\widehat{\overline{Y}}_R] &\approx \frac{1-f}{n} \left(S_y^2 + R^2 S_x^2 - 2RS_{xy} \right) \\ \widehat{V}[\widehat{\overline{Y}}_R] &= \frac{1-f}{n} \left(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R} s_{xy} \right) \end{aligned}$$

where S_y^2, S_x^2, S_{xy} are the population quasivariances and quasicovariance of the Y values, and s_y^2, s_x^2, s_{xy} the corresponding over the sample. Also, we denote $f = n/N$.

An alternative and more formal approach is the **predictive approach**, based on the existence of the following superpopulation model,

$$\boxed{\begin{array}{l} Y_i = \beta X_i + \varepsilon_i \\ \\ E_s[\varepsilon_i] = 0 \\ V_s[\varepsilon_i] = \sigma^2 v(X_i) \\ E_s[\varepsilon_i \varepsilon_j] = 0, \ i \neq j \end{array}}$$

where $v(\cdot)$ is a known function. We suppose that X only takes positive values. Under this model, and if $v(x) = x$, we obtain the estimator,

$$\widehat{\overline{Y}}_R = \frac{\sum_{i \in s} Y_i/\pi_i}{\sum_{i \in s} X_i/\pi_i} \overline{X}$$

particular case of Sánchez-Crespo (1980).

The properties of this estimator depends on the sampling design. If the sample, s, is obtained using a SRS(N,n), then, the estimator is,

$$\widehat{\overline{Y}}_R = \frac{\overline{y}}{\overline{x}} \overline{X}$$

This estimator is biased, since,

$$B[\widehat{\overline{Y}}_R] = E[\widehat{\overline{Y}}_R] - \overline{Y} = \overline{X} E\left[\frac{\overline{y}}{\overline{x}}\right] - E[\overline{y}] = E[\overline{x}] E\left[\frac{\overline{y}}{\overline{x}}\right] - E[\overline{y}] = -\operatorname{Cov}\left[\overline{x}, \frac{\overline{y}}{\overline{x}}\right]$$

and defining the quantities,

$$\begin{aligned} B_k[\widehat{\overline{Y}}_R] &= \overline{Y} E\left[\sum_{i=0}^{2k-1} (-\delta\overline{x})^i (\delta\overline{y} - \delta\overline{x})\right] \\ \mathrm{ECM}_k[\widehat{\overline{Y}}_R] &= \overline{Y}^2 E\left[(\delta\overline{y} - \delta\overline{x})^2 \sum_{i=0}^{2k-2} (i+1)(-\delta\overline{x})^i\right] \end{aligned}$$

we have the following result, about the bias and the mean square error,

$$\begin{aligned} \left| B[\widehat{\overline{Y}}_R] - B_k[\widehat{\overline{Y}}_R] \right| &\leq O(n^{-(k+1)}) \\ \left| \mathrm{ECM}[\widehat{\overline{Y}}_R] - \mathrm{ECM}_k[\widehat{\overline{Y}}_R] \right| &\leq O(n^{-(k+1)}) \end{aligned}$$

If $k=1$, we obtain the first order approximations,

$$\begin{aligned} B[\widehat{\overline{Y}}_R] &= \frac{1-f}{n}\left(\frac{\overline{Y}}{\overline{X}^2}S_x^2 - \frac{1}{\overline{X}}S_{xy}\right) + O(n^{-2}) = O(n^{-1}) \\ \mathrm{ECM}[\widehat{\overline{Y}}_R] &= \frac{1-f}{n}\left(S_y^2 + \frac{\overline{Y}^2}{\overline{X}^2}S_x^2 - 2\frac{\overline{Y}}{\overline{X}}S_{xy}\right) + O(n^{-2}) = O(n^{-1}) \\ V[\widehat{\overline{Y}}_R] &= \frac{1-f}{n}\left(S_y^2 + \frac{\overline{Y}^2}{\overline{X}^2}S_x^2 - 2\frac{\overline{Y}}{\overline{X}}S_{xy}\right) + O(n^{-2}) = O(n^{-1}) \end{aligned}$$

Thus, the bias and the mean square error can be controlled, increasing the sample size.

If we use a sampling design with first order inclusion probabilities proportional to size, X, we obtain the unbiased estimation,

$$\widehat{\overline{Y}}_R = \frac{1}{n}\sum_{i\in s}\frac{Y_i}{X_i}\overline{X}$$

with variance (for fixed sample size),

$$\begin{aligned} V[\widehat{\overline{Y}}_R] &= -\frac{\overline{X}^2}{2n^2}\sum_{i,j\in U}(\pi_{ij}-\pi_i\pi_j)\left(\frac{Y_i}{X_i}-\frac{Y_j}{X_j}\right)^2 \\ \widehat{V}[\widehat{\overline{Y}}_R] &= -\frac{\overline{X}^2}{2n^2}\sum_{i,j\in s}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{Y_i}{X_i}-\frac{Y_j}{X_j}\right)^2 \end{aligned}$$

that is to say, the variance decreases if the Y and X variables are proportional or nearly proportional.

Other unbiased strategies are obtained combining the ratio estimator,

$$\widehat{\overline{Y}}_R = \frac{\overline{y}}{\overline{x}}\overline{X}$$

with the sampling design generated by the Lahiri-Midzuno scheme.

The Harley-Ross estimator,

$$\widehat{\overline{Y}}_{HR} = \overline{z}\overline{X} + \frac{n(N-1)}{N(n-1)}(\overline{y}-\overline{zx}), \qquad Z_i = Y_i/X_i$$

with the $\mathrm{SRS}(N,n)$ design.

The Mickey's estimator with the SRS(N,n) design, and others related by Ruiz and Santos (1989).

Also, is possible to obtain almost unbiased strategies, that is to say, with bias $O(1/n^2)$, for example,

- The De Pascual's estimator,

$$\widehat{\overline{Y}}_{DP} = \widehat{\overline{Y}}_R + \frac{\overline{y} - \overline{zx}}{n-1} \qquad \text{with } Z_i = \frac{Y_i}{X_i}$$

- The Beale's estimator,

$$\widehat{\overline{Y}}_B = \widehat{\overline{Y}}_R \left[1 + \left(\frac{1}{n} - \frac{1}{N}\right) \frac{s_{xy}}{\overline{xy}}\right] \left[1 + \left(\frac{1}{n} - \frac{1}{N}\right) \frac{s_x^2}{\overline{x^2}}\right]^{-1}$$

- The Tin's estimator,

$$\widehat{\overline{Y}}_T = \widehat{\overline{Y}}_R \left[1 - \left(\frac{1}{n} - \frac{1}{N}\right) \left(\frac{s_x^2}{\overline{x^2}} - \frac{s_{xy}}{\overline{xy}}\right)\right]$$

in combination with the SRS(N,n) design. And also, the estimator obtained applying the jackknife technique to the ratio estimator, $\widehat{\overline{Y}}_R = (\overline{y}/\overline{x})\overline{X}$.

To finish this review, we study the multivariate ratio estimator, the ratio estimator in stratified sampling and the optimality of the ratio estimator.